Research & Methods
ISSN 1234-9224 Vol. 21 (1, 2012): 5–29
Institute of Philosophy and Sociology
Polish Academy of Sciences, Warsaw
www.ifispan.waw.pl
e-mail: publish@ifispan.waw.pl

Measuring Attitudes toward Immigration in Europe: The Cross–cultural Validity of the ESS Immigration Scales

Bart Meuleman
Centre for Sociological Research (CeSO), University of Leuven, Belgium

Jaak Billiet
Centre for Sociological Research (CeSO), University of Leuven, Belgium

Equivalence of measurement scales is a crucial prerequisite for making valid cross-cultural comparisons, as cultural differences in the interpretation of indicators could result in misleading conclusions. In this paper, we empirically assess the cross-national measurement equivalence of four scales that are included in the European Social Survey, round 1 (2002–03). These four scales, referring to various aspects of attitudes toward immigration, are: (1) opposition against new immigration into the country (REJECT), (2) support for imposing conditions to immigration (CONDITION), (3) perceived economic threat (ECOTHREAT) and (4) perceived cultural threat (CULTHREAT).

To test for measurement equivalence, we make use of multi-group confirmatory factor analysis (MGCFA). In this approach, a distinction is made between configural (equal factor structures), metric (equal factor loadings) and scalar (equal item intercepts) equivalence. A step-by-step strategy to test for these distinctive levels of equivalence is explained in a detailed manner.

Our results show that the degree of cross-cultural equivalence differs quite strongly from one scale to another. In the case of the REJECT-scale, the number of violated equality constraints is limited, and partial scalar equivalence is found to hold for all countries. The other measurement scales are cross-culturally less robust, and comparability is only guaranteed for subsets of countries.

Key words: measurement equivalence; multigroup CFA; anti-immigration attitudes; perceived threat; European Social Survey

INTRODUCTION

Thanks to the increasing availability of cross-national survey data, a steadily growing number of studies focuses on international comparisons of value and attitude patterns. This tendency is remarkably outspoken in the field of anti-immigration attitudes and perceived ethnic threat (to cite a few recent examples of such studies: Bail 2008; Davidov & Meuleman 2012; Kunovich 2004; Meuleman, Davidov & Billiet 2009; Schlueter & Scheepers 2010; Schneider 2008; Semyonov, Raijman & Gorodzeisky 2006; Sides & Citrin 2007; Strabac & Listhaug 2008; for a review of recent developments in the field, see Ceobanu & Escandell 2010).

However, the cross-cultural comparison of abstract psychological constructs –such as anti-immigration attitudes– brings along methodological problems that do not present themselves in single nation research. Among many others, the comparability of measurements is an important issue: It is not sure whether a given set of indicators taps into the same concept in different countries, since cross-cultural differences in the interpretation of items might exist. And even if the same concept is measured across countries, it is far from guaranteed that this concept is measured on the same measurement scale. Thus, before meaningful comparisons can be made, equivalence of the measurements has to be assessed. Here, measurement equivalence refers to the question '*whether or not, under different conditions of observing and studying phenomena, measurement operations yield measures of the same attribute*' (Horn & McArdle 1992: 117).

In this paper, we test the critical assumption of measurement equivalence for several scales measuring attitudes towards immigration included in the European Social Survey (ESS). For this purpose, a multi-group confirmatory factor analysis (MGCFA) approach is used (Jöreskog 1971; Steenkamp & Baumgartner 1998). The contribution of this study is twofold. First, the cross-cultural validation of these very frequently used scales provides valuable insights for the numerous researchers working with ESS data and presents useful information for future questionnaire design. Second, besides providing a test of these specific scales, our analysis also serves didactical purposes, as we illustrate how measurement equivalence can be tested more in general. By discussing several often neglected but nevertheless crucial practical issues explicitly (such as the order of the tests and criteria to determine whether inequivalence is present), we hope to make this toolkit more accessible for applied researchers.

This paper sets out by explaining a concrete research strategy to assess measurement equivalence using MGCFA. Second, we give a brief presentation of the ESS scales concerning attitudes towards immigration and perceived threat that are analyzed in this study. In section 3, finally, the results of the measurement

equivalence tests are presented. A conclusion and discussion section complete the paper.

1. THE MGCFA APPROACH TO MEASUREMENT EQUIVALENCE

Several post-survey techniques have been proposed to test the assumption of measurement equivalence, such as confirmatory factor analysis, latent class analysis, item response theory or multi-dimensional scaling (for an overview of techniques, see Van de Vijver & Leung 1997; Johnson 1998; for a comparison, see Kankaras, Vermunt & Moors 2011). These techniques share the central principle that measurements are considered as equivalent when the relations between the indicators and the traits these indicators are measuring are invariant across countries (Reise et al. 1993). In this study, we opt for the multi-group confirmatory factor analytic (MGCFA) approach. This versatile tool is probably the most often used and therefore probably the most developed technique to test for measurement equivalence (Drasgow & Kanfer 1985; Byrne et al. 1989; Marsh 1994; Steenkamp & Baumgartner 1998; Cheung & Rensvold 1999; Billiet 2003).

1.1 The MGCFA model

Confirmatory factor analysis (CFA) starts from a measurement model, in which latent variables are indicated by observed indicators (Brown 2006; Bollen 1989). In CFA, observed responses x_i (i = 1,..,p) are written as linear functions of latent variables ξ_j (j = 1,…,m).

$$\mathbf{x} = \boldsymbol{\tau} + \boldsymbol{\Lambda}\boldsymbol{\xi} + \boldsymbol{\delta} \tag{1}$$

In expression (1), $\mathbf{x}$ refers to a $p \times 1$ vector containing the observed responses. This vector is modeled as the sum of three components. $\boldsymbol{\Lambda}\boldsymbol{\xi}$ is the product of a $p \times m$ matrix containing the factor loadings ($\boldsymbol{\Lambda}$) and a $m \times 1$ vector with the latent variable scores. The factor loadings can be seen as the slopes of a regression of x_i on ξ_j, while $\boldsymbol{\tau}$ is a $p \times 1$ vector with the intercepts of the functions. These intercepts refer to the expected value of the observed indicators when the latent variable score is equal to zero. Finally, $\boldsymbol{\delta}$ is a $p \times 1$ vector containing stochastic error terms that are assumed to follow a multivariate normal distribution and to have expected value 0. When correctly identified, this measurement model implies the following mean structure $\boldsymbol{\mu}$ and covariance structure $\boldsymbol{\Sigma}$:

$$\boldsymbol{\mu} = \boldsymbol{\tau} + \boldsymbol{\Lambda}\boldsymbol{\kappa} \tag{2}$$

$$\boldsymbol{\Sigma} = \boldsymbol{\Lambda}\boldsymbol{\Phi}\boldsymbol{\Lambda}' + \boldsymbol{\Theta} \tag{3}$$

where $\boldsymbol{\mu}$ equals a $p \times 1$ vector with observed item means and $\boldsymbol{\kappa}$ a $m \times 1$ vector with means of latent variables ξ_j; $\boldsymbol{\Sigma}$ is the $p \times p$ covariance matrix of the observed indicators, $\boldsymbol{\Phi}$ a $m \times m$ covariance matrix of the latent variables and $\boldsymbol{\Theta}$ a $p \times p$ matrix with the error (co)variances (Steenkamp & Baumgartner 1998). Comparing these implied mean and covariance structures with their observed counterparts makes it possible to assess how good the measurement model fits the data. Formally, this comparison can be made using the chi-square difference test (which is known to be very sensitive for large sample sizes, however) or several alternative fit indices, such as the RMSEA or CFI (Browne & Cudeck 1992; Bentler 1990).

In order to be useful for measurement equivalence testing, the factor analytic model described above has to be extended to a multi-group setting (Jöreskog 1971). Concretely, this means that the system of equations of (1) is estimated separately but simultaneously for different groups g (g = 1…G) of respondents.

$$\mathbf{x}^g = \boldsymbol{\tau}^g + \boldsymbol{\Lambda}^g\boldsymbol{\xi} + \boldsymbol{\delta}^g \tag{4}$$

In this case, the groups are obviously formed by inhabitants of different countries. Measurement equivalence is then assessed by comparing certain parameter estimates over groups (countries).

This MGCFA model present above presupposes continuous indicators that follow a multivariate normal distribution. Various extensions of the MGCFA model have been presented to deal with ordered-categorical indicators (Millsap & Yun-Tein 2004; Davidov et al. 2011). Here, we follow the approach implemented in the LISREL program. The categorical indicators are assumed to be discretized versions of underlying latent response variables. Concretely, an observed ordered-categorical indicator with $c+1$ categories is obtained by partitioning the latent response variable along c thresholds. Jöreskog (1990) proposes a procedure to estimate polychoric correlation and asymptotic covariance matrices reflecting the relations between the underlying response variables. These matrices are then analyzed using a weighted least squares approach.

1.2 Levels of measurement equivalence

Steenkamp & Baumgartner (1998) discern various levels of measurement equivalence that can be assessed within the MGCFA framework. These levels are ordered hierarchically in the sense that higher equivalence levels presuppose lower ones. Higher equivalence levels are harder to obtain as they provide a stronger test of cross-cultural equivalence, but also allow a more extended form of cross-cultural comparison.

Configural equivalence is the basic level of equivalence in the Steenkamp & Baumgartner (1998) scheme. Configural equivalence means that the measurement model for the latent concept has the same factor structure across cultural groups. In other words, configural equivalence implies that the items in the measurement instrument exhibit the same configuration of salient and nonsalient factor loadings across countries. However, the strength of the factor loadings can differ across countries, as no restrictions are placed on the magnitude of these parameters (Steenkamp & Baumgartner 1998: 80). The conditions for configural equivalence can be written formally as follows:

- if λ_{ij}^{g} is close to 0, then λ_{ij}^{h} is close to 0 for $g,h = 1 \ldots G$ (where superscripts g and h refer to two different groups) (5)
- if λ_{ij}^{g} is not close to 0, then λ_{ij}^{h} is not close to 0 for $g,h = 1 \ldots G$; $g \neq h$

where λ_{ij}^{g} is the factor loading of item x_i on latent variable ξ_j for group g. Generally, this basic level of measurement equivalence is relatively easy to reach. The other side of the coin is that configural equivalence does not guarantee any cross-cultural score comparability. Configural equivalence instead means that the latent concepts can be meaningfully discussed in all countries. Configural equivalence is often used as a baseline for further equivalence testing (Vandenberg & Lance 2000).

A second and higher level of equivalence is called *metric equivalence* (Steenkamp & Baumgartner 1998), although it has also been referred to as construct equivalence (Van de Vijver & Leung 1997). Operationally, measurement equivalence presupposes that the factor loadings in the measurement model are invariant over groups. Formally, this can be written as follows:

$$\mathbf{\Lambda}^1 = \mathbf{\Lambda}^2 = \ldots = \mathbf{\Lambda}^G \tag{6}$$

Metric equivalence implies the cross-cultural equality of the intervals of the scale on which the latent concept is measured. In other words: An increase of 1 unit on the measurement scale has the same meaning in country A as in country B. However, latent variable scores can still be uniformly biased upward or downward. Because of this possibility of additive bias, metric equivalence still does not lead to full score comparability. Nevertheless, metric equivalence is highly relevant because it makes comparison of difference scores (i.e. mean-corrected scores) across countries possible (Steenkamp & Baumgartner 1998: 80). Since regression coefficients and covariances are based on such difference scores, metric equivalence guarantees their comparability.

An even stronger test for measurement equivalence is *scalar equivalence.* Within the MGCFA framework, scalar equivalence can be defined as the equality of intercept parameters over groups:

$$\boldsymbol{\tau}^1 = \boldsymbol{\tau}^2 = \ldots = \boldsymbol{\tau}^G \tag{7}$$

Concretely this means that all observed mean differences in the items must be conveyed through mean differences in the latent factor. In other words, respondents from different countries with the same value on the latent factor should exhibit the same expected score on the observed indicators. Scalar equivalence implies that the measurement scales do not only have the same intervals, but also share origins. This makes it possible to compare raw scores in a valid way, which is a prerequisite for country-mean comparisons. Variables that are analyzed by means of multilevel modeling also need to be measured in an at least scalar equivalent way. After all, multilevel models rely on country-specific means of the dependent variable to estimate the random intercept variance.

Apart from the three forms of equivalence described above, Steenkamp & Baumgartner (1998) mention even higher levels of equivalence, such as factor variance equivalence and error variance equivalence. Yet, since cross-cultural researchers are not frequently interested in comparing variances of latent traits or measurement errors of indicators, we leave these levels of equivalence aside.

However, invariance of the parameters for all items is not necessary in order for substantive analyses to be meaningful. Byrne et al. (1989) argue that that valid comparisons are also possible under the condition of *partial equivalence*. Partial equivalence requires that the measurement parameters of at least two items per construct are identical across all groups. After all, fixing two items, namely the marker item for which the loading is fixed at unity to identify the model and one other item, is sufficient to determine the metric of the scale (these two items will be called the calibration items in the remainder of this article). Thus, setting equivalence constraints free for some (but not all) items can control for the measurement inequivalence caused by a limited number of violations of the equivalence requirements (Vandenberg & Lance 2000: 37). This idea is also supported by Steenkamp and Baumgartner (1998).

1.3 A concrete strategy to test for measurement equivalence

In the literature, there is considerable agreement on the specific hypotheses – see equations (5), (6) and (7) – that need to be tested to assess various levels of measurement equivalence. There is far less consensus, however, regarding the optimal procedures to test these hypothesis in practice. First, there exists a wide

variety of opinions on the order in which the different test are carried out best (for a review of practices, see Vandenberg & Lance 2000). Second, confusion has arisen regarding appropriate criteria for deciding whether the equivalence hypotheses are violated or not (Meuleman 2012). These important issues are rarely discussed explicitly, leading to a lack of transparency in the field of equivalence testing. By proposing a concrete strategy to test for measurement equivalence, we aim at making this technique more accessible for applied researchers.

Regarding the order of the tests, we propose a bottom-up logic, starting at the lowest level of equivalence (see also Steenkamp & Baumgartner 1998; for an example of a top-down strategy, see Horn & McArdle 1992). In a first step, configural equivalence is tested for. If this basic level of equivalence is violated, no meaningful cross-country comparisons are possible, and further analysis has to be precluded.

In the case configural equivalence holds, it is possible to move up the equivalence ladder and to estimate a full metric equivalence model (i.e. equality of all factor loadings). Subsequently, it should be tested whether the model can be improved by setting one or more of the factor loadings free across countries. This confronts us with the question, of course, on what basis one could decide whether relaxing a parameter constraint leads to a meaningful model improvement. A typical approach would consist of inspecting modification indices (MIs; these are in fact chi²-test statistics with one degree of freedom) for the constrained model. Significant MIs are then indicative of model misfit, and the parameter constraints they refer to should be relaxed. However, various authors warn against relying on statistical criteria alone, because due to the large sample sizes often used, even negligible differences between groups can become significant (Rensvold & Cheung 1998; Saris et al 2009; Vandenberg & Lance 2000). Saris et al. (1987) argue that setting free parameter constraints is only relevant when this leads to substantive parameter changes (as indicated by the expected parameter change [EPC]).

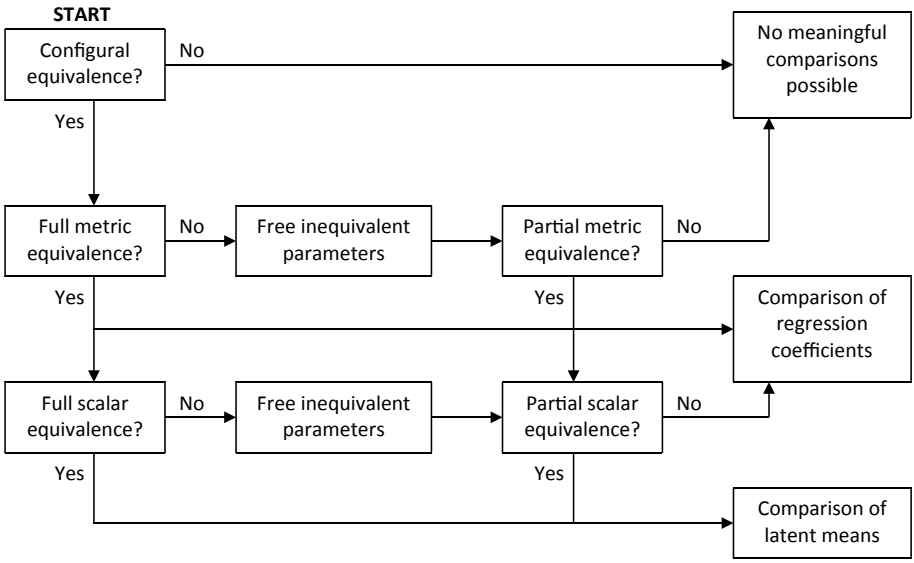
In order to avoid over-fitting and a data driven approach, we propose a strategy drawing on this idea by Saris et al. (1987). After estimating the full metric equivalence model, the factor loading with the highest MI should be identified. If this MI is strongly significant ($p<.0001$; this strict alpha-level implies a Bonferroni-type correction for the fact that multiple tests are conducted at the same time – Rensvold & Cheung 1998), and if the associated EPC is substantively relevant, the equivalence constraint should be considered as untenable and therefore be relaxed. Subsequently, the second-highest MI should be looked into. This iterative model-fitting process should be repeated until no possibilities to improve the model substantially are left.

If, once that no further possibilities to improve the model are left, at least partial metric equivalence is present (i.e. when the factor loadings for at least two items

are equal across groups), regression coefficients can be compared in a valid way across countries. If not, there is still a possibility to search for subsets of countries for which (partial) metric equivalence does hold. Once (partial) metric equivalence is evidenced, we proceed in a similar way to test for (partial) scalar equivalence (i.e. looking for intercepts with high MIs and EPCs).

The flowchart in Figure 1 gives a graphical representation of this procedure.

Figure 1 Flowchart of the proposed procedure for equivalence testing



2. PRESENTATION OF THE ESS IMMIGRATION SCALES

The strategy explained in the previous section will now be used to test the cross-cultural comparability of scales measuring attitudes toward immigration included in the ESS. Because the first round of the ESS (2002/03) contains a complete module covering various aspects of attitudes toward immigration, this dataset will be analyzed. The results bear relevance for other ESS rounds as well, as several items were retained in the core module of ESS.[1] The countries[2] in this study (and the effective sample sizes[3]) are: Austria (AT) (1,800), Belgium (BE) (1,683), Czech Republic (CZ) (1,156), Denmark (DK) (1,333), Finland (FI) (1,890), France (FR) (1,316), Germany (DE) (2,625), Great Britain (GB) (1,837), Greece (GR) (2,206), Hungary (HU) (1,348), Ireland (IE) (1,816), Italy (IT) (1,050), Luxemburg (LU)

(841), the Netherlands (NL) (1,919), Norway (NO) (2,173), Poland (PL) (1,764), Portugal (PT) (1,053), Slovenia (SI) (1,364), Spain (ES) (1,268), Sweden (SE) (1,754), and Switzerland (CH) (1,659).

We restrict ourselves to a series of items (see Table 1 for question wording and answer scales) that, based on theoretical arguments, can be grouped into four constructs.[4] Two constructs refer to preferences for certain immigration policies, two others measure particular forms of perceived ethnic threat. The first theoretical construct, REJECT, measures opposition against allowing new immigrants into the country and thus indicates preferences for a restrictive immigration policy. The items measuring this concept (d4-d8) inquire whether respondents prefer their country to allow many or few immigrants of specific groups (e.g. 'immigrants of a different race or ethnic group', 'immigrants from the poorer countries in Europe'). 4-point scales (1 – allow many, 2 – allow some, 3 – allow a few, 4 – allow none) are used to register the answers of the respondents. A second immigration policy-related construct, CONDITION (measured by items d10-d12 and d16), refers to preferences for imposing conditions on immigration flows. Respondents were asked how important they think certain qualifications (such as education qualifications, work skills or language knowledge) are in the decision whether an immigrant is granted entrance to the country.

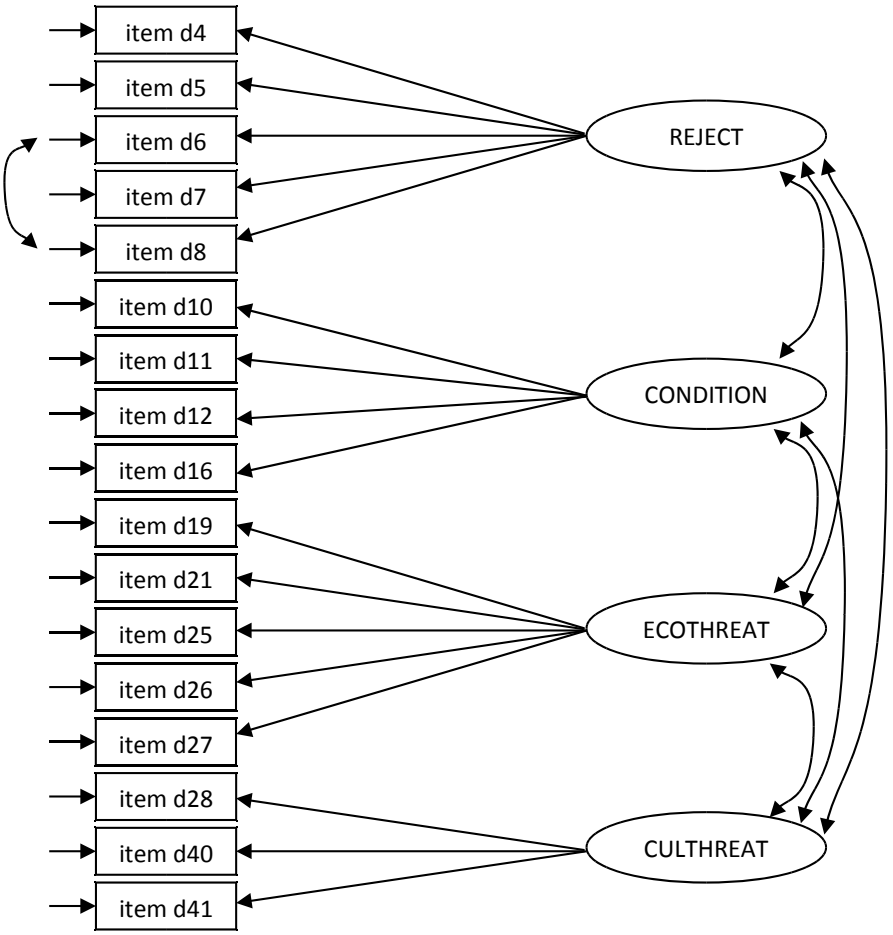
We distinguish between two types of perceived ethnic threat. Economic threat (ECOTHREAT – items d19, d21, d25, d26 and d27) refers to fears that the own social group has to compete with immigrants for scarce material goods (such as jobs or resources of the welfare state). Cultural threat (CULTHREAT – items d28, d40 and d41), on the other hand, relates to intergroup competition for symbolic rather than material goods. It is the perception that immigrants adhering to different cultural traditions pose a threat to the own worldview, that is believed to be morally right (Stephan et al. 1998).

By means of MGCFA, we assessed the measurement quality of these four scales. We estimated a multigroup model for the 21 countries containing the four latent variables (without cross-group equality constraints). The model was estimated with LISREL 8.7 (Jöreskog & Sörbom 1993). Because all items are measured on ordinal scales and most items have a strongly skewed distribution, we decided to use a weighted least squares (WLS) estimation procedure, in which polychoric correlations and asymptotic covariance matrices are used as input rather than regular covariance matrices (Jöreskog 1990). An error correlation is tolerated between items d6 and d8. This error correlation is theoretically justified since both items have an element in common that is not captured by the latent factor, namely they both refer to the very specific group of immigrants from richer countries. Figure 2 gives a graphical representation of this model.

Table 1 Question wording of the ESS immigration items used in this study

	Question wording	Answer categories
REJECT	To what extent do you think [country] should allow people …	
	D4. ... of the same race or ethnic group from most [*country*] people to come and live here?	1 (many), 2 (some), 3 (a few), 4 (none)
	D5. ... of a different race or ethnic group from most [*country*] people to come and live here?	
	D6. … from the richer countries in Europe?	
	D7. ... from the poorer countries in Europe to come and live here?	
	D8. ... from the richer countries outside Europe to come and live here?	
CONDITION	Please tell me how important you think each of these things should be in deciding whether someone born, brought up and living outside [country] should be able to come and live here. How important should it be for them to …	0 (extremely unimportant) to 10 (extremely important)
	D10. ... have good educational qualifications?	
	D11. ... have close family living here?	
	D12. ... be able to speak [*country language*]?	
	D16. ... have work skills that [*country*] needs?	
ECOTHREAT	D. 19 People who come to live and work here generally harm the economic prospects of the poor more than the rich	1 (agree strongly) to 5 (disagree strongly)
	D21. If people who have come to live and work here are unemployed for a long period, they should be made to leave.	
	D25. Would you say that people who come to live here generally take jobs away from workers in [*country*], or generally help to create new jobs?	0 (take jobs away) to 10 (create new jobs)
	D26. Most people come to live here work and pay taxes. They also use health and welfare services. On balance, do you think people who come here take out more than they put in or put in more than they take out?	0 (generally take out more) to 10 (generally put in more)
	D27. Would you say that it is generally bad or good for [*country*] economy that people come to live here from other countries?	0 (bad for the economy) to 10 (good for the economy)
CULTHREAT	D28. Would you say that [*country*] cultural life is generally undermined or enriched by people coming to live here from other countries?	0 (cultural life undermined) to 10 (cultural life enriched)
	D40. It is better for a country if almost everyone shares the same customs and traditions.	1 (agree strongly) to 5 (disagree strongly)
	D41. It is better for a country if there are a variety of different religions.	

Figure 2 Graphical representation of the configurally equivalent model



$\chi^2 = 12{,}624.35$ df = 2352 RMSEA = 0.052 CFI = 0.99

The model confirms that the theoretical constructs are measured in a sufficiently valid and reliable way. Virtually all standardized factor loadings are larger than 0.40,[5] and no cross-loadings had to be tolerated. With 2352 degrees of freedom (112 in each of the 21 countries), the chi-square value amounts to 12,624.35. The Root Mean Square Error of Approximation (RMSEA) is equal to 0.052 and the Comparative Fit Index (CFI) is 0.99. Since RMSEA is not much higher than 0.05 and CFI is sufficiently close to 1 (Byrne 1998; Hu & Bentler 1999), the formulated measurement model gives an acceptable description of the data.

The four substantive factors that are distinguished in the model are not unrelated, but instead correlate quite strongly.[6] Those who perceive a strong economic and cultural threat and are more in favour of a restrictive immigration policy and imposing conditions to immigration.

The main conclusion from this first measurement model is that four postulated theoretical constructs are measured adequately in the countries under study. Implicitly, also configural equivalence of the concepts was shown, since the measurement model has the same structure of salient and non-salient factor loadings across 21 countries.

3. CROSS-CULTURAL COMPARABILITY OF THE ESS IMMIGRATION SCALES

Configural equivalence does not yet guarantee that valid comparisons across countries are possible, however. In this third section, we assess both metric and scalar equivalence for the four constructs. The tests are performed for each scale separately as estimating a measurement model for all scales simultaneously would be computationally too demanding.

3.1 Metric and scalar equivalence tests for REJECT

Table 2 summarizes the model fitting process for the REJECT-scale. Every row of the table represents a new model (often after setting a parameter free). The table gives a short description of the model, as well as fit indices (chi², RMSEA, CFI, the difference in chi² compared to the previous model) and the EPC value for the parameter that is set free in the model. At the bottom of the table, the common estimates for factor loadings and intercepts can be found.

We start from configural equivalence, which is used as a baseline model. This model without cross-country equality constraints has an acceptable fit. The RMSEA (0.046) falls well below common cut-off points (Byrne, 1998; Hu & Bentler, 1999), and the CFI is equal to one.

In a second step, full metric equivalence is tested for by setting all factor loadings equal across countries. Judging by the RMSEA, this fully metric equivalent model fits the data even better. However, the primary goal of this analysis is not to evaluate the overall model fit, but rather to test whether the equality constraints on the factor loadings are tenable. Modification indices suggest that the model can be improved by freeing the equality constraint for the loading of the first item in Hungary. The standardized EPC (-0.33) indicates that deleting this constraint would indeed cause a substantial modification in the estimate of the factor loading. Therefore, the model was re-estimated without this equality constraint. Giving up one degree of freedom causes the chi-square value to drop by more than 120. The modified model clearly gives an even better description of the data. After this

modification, no possibilities for further substantial model improvement were left. Since a deviation for only one factor loading was detected, and the factor loadings for the four other items are equal across all groups, partial metric equivalence is evidenced for the scale REJECT. This is a condition for valid comparisons of regression coefficients.

In a next step, the mean structure of the items is added to the model and scalar equivalence is tested by setting all intercepts equal across countries. The factor loading that was set free during metric equivalence testing is still estimated as a free parameter. This time, the model could be improved substantially five times by freeing a constrained intercept. The final model has a good overall model fit. Two items (d5 and d7) have invariant measurement parameters (loadings and intercepts) over all countries, and therefore partial scalar equivalence holds (Byrne et al. 1989; Steenkamp & Baumgartner 1998). This leads to the conclusions that means on the latent variable REJECT can be validly compared across all 21 ESS round 1 countries.

Table 2 Equivalence tests for REJECT – N = 33,855

Model specifications		χ^2	df	RMSEA	CFI	$\Delta\chi^2$	EPC
M0	Configural equivalence	371.10	84	0.046	1.00	- -	- -
M1	Full metric equivalence	703.03	164	0.045	1.00	- -	- -
M2	+ λ_{d4}^{HU} free	579.39	163	0.040	1.00	123.64	-0.33
M3	Scalar equivalence	1076.92	243	0.046	1.00	- -	- -
M4	+ τ_{d4}^{HU} free	988.93	242	0.044	1.00	87.99	-0.33
M5	+ τ_{d8}^{CZ} free	948.93	241	0.043	1.00	40.00	-0.42
M6	+ τ_{d6}^{GR} free	902.69	240	0.041	1.00	46.24	-0.27
M7	+ τ_{d8}^{PL} free	870.08	239	0.040	1.00	32.61	-0.29
M8	+ τ_{d6}^{DK} free	822.32	238	0.039	1.00	47.76	-0.28

Common solution for final model

λ_{d4}	$^*\lambda_{d5}$	λ_{d6}	λ_{d7}	λ_{d8}	τ_{d4}	$^*\tau_{d5}$	τ_{d6}	τ_{d7}	τ_{d8}
0.98	1.00	0.85	0.99	0.86	0.14	0.14	0.14	0.13	0.16

* Marker item (loading fixed to 1 to identify the model)

The most notable inequivalence is found for Hungary. We agree with Poortinga (1989) that a lack of equivalence is not just a source of error, but can be considered as a source of useful information as well. The fact that certain items function differently across countries might reveal important information on cross-cultural

differences. The lower factor loading for Hungary, for example, indicates that the first item on the scale is not as strongly connected to the whole scale as in other countries. Apparently, attitudes towards immigrants of the same ethnic group are rather detached from attitudes toward other immigrant groups in Hungary, while these attitudes are more strongly intertwined in other countries. Also, the intercept of the first item was found to be lower in Hungary. This means that Hungarians have less restrictive attitudes toward immigrants of the same ethnic group than what is expected based on their score on the latent variable REJECT. A possible explanation for this differential functioning of item d4 might be sought in the specific ethnic and immigration context of Hungarian society. A large number of ethnic Hungarians living in the neighboring countries. This specific setting apparently colors the way in which Hungarian citizens interpret this particular item.

Furthermore, inhabitants of Czech Republic and Poland are, compared to their general stance on immigration policies, relatively open toward the group of immigrants from the richer countries outside Europe. The Greeks and Danish are relatively prepared to allow immigrants from richer countries in Europe. An explanation for these deviations is not immediately clear.

3.2 Metric and scalar equivalence tests for CONDITION

Metric equivalence tests for the CONDITION-scale indicate that strong cross-country differences exist with respect to strength of the factor loading of d11. This item inquires whether respondents are of the opinion that having close family in the country of destination should be an important condition for allowing immigrants. In Spain, Greece, Hungary, Poland and Portugal, this item loads very strongly (>.70) on latent factor CONDITION, while in Germany, Denmark, the Netherlands, and Norway low factor loadings (<.40) are retrieved. In the former countries, having close family is seen as forming a part of the whole of conditions for immigration, where in the latter countries, having family is far less related to the importance of imposing conditions. These cross-cultural differences are related to the role the family plays in different societies. It is no coincidence that the factor loading is strong in more traditional or catholic Southern and Eastern European countries, whereas low loadings are found in Northern, more individualized societies. Because of the large amount of violations of metric equivalence, item d11 was dropped from further analysis.

After dropping item d11, the fit of the configural equivalence model is no longer informative. Since there are only three items, the model is just identified and therefore has a perfect fit. To test for metric equivalence, factor loadings were constrained to be equal across countries in a second step. Five deviations from full metric equivalence were found. Once the possibilities for model improvement

were exhausted, the model had a good overall fit. Since deviations are found for all items, partial metric equivalence does not hold for all countries. Inevitably, some countries will have to be excluded from substantive analysis. Depending on the two items that are chosen to determine the metric of the scale (i.e. calibration of the scale), various sets of countries for which measurements are partially equivalent can be formed. If items d10 and d12 are chosen, for example, then Finland, France and Great Britain cannot be included in valid cross-country comparisons of e.g. regression coefficients. After all, each of these three country has a deviating parameter for one of the calibration items.

Table 3 Equivalence tests for CONDITION – N = 33,855

Model specifications		χ^2	Df	RMSEA	CFI	$\Delta\chi^2$	EPC
M0	Configural equivalence	0.00	0	0	1	- -	- -
M1	Full metric invariance	278.47	40	0.061	0.99	- -	- -
M2	λ_{d16}^{DK} free	235.84	39	0.056	0.99	42.63	0.16
M3	λ_{d12}^{FR} free	197.92	38	0.051	0.99	37.92	0.18
M4	λ_{d12}^{GB} free	173.66	37	0.048	0.99	24.26	0.12
M5	λ_{d16}^{IE} free	142.49	36	0.043	1	31.17	-0.16
M6	λ_{d10}^{FI} free	122.30	35	0.039	1	20.19	-0.10
M7	Scalar equivalence	1258.83	75	0.099	0.95	- -	- -
M8	τ_{d12}^{LU} free	997.59	74	0.088	0.96	261.24	0.88
M9	τ_{d12}^{NO} free	763.46	73	0.077	0.97	234.13	0.5
M10	τ_{d12}^{DE} free	702.20	72	0.074	0.97	61.26	0.28
M11	τ_{d16}^{PT} free	629.33	71	0.07	0.98	72.87	0.4
M12	τ_{d12}^{NL} free	571.45	70	0.067	0.98	57.88	0.25
M13	τ_{d12}^{SI} free	512.33	69	0.063	0.98	59.12	-0.31
M14	τ_{d16}^{CZ} free	477.42	68	0.061	0.98	34.91	0.35
M15	τ_{d16}^{HU} free	433.85	67	0.058	0.98	43.57	0.30
M16	τ_{d12}^{GB} free	399.64	66	0.056	0.99	34.21	0.26
M17	τ_{d12}^{FR} free	364.08	65	0.053	0.99	35.56	0.34
M18	λ_{d16}^{IT} free	337.48	64	0.051	0.99	26.60	-0.20

Common solution for the final model

$^*\lambda_{d10}$	λ_{d12}	λ_{d16}	$^*\tau_{d10}$	τ_{d12}	τ_{d16}
1.00	0.88	0.98	0.26	0.09	0.20

* Marker item (loading fixed to 1 to identify the model)

Not surprisingly, scalar equivalence is even more problematic than metric equivalence. After imposing cross-cultural invariance of intercepts, 12 additional model improvements were possible. Seven of these improvements relate to the item on speaking the country's language as a condition for immigration (d12). Luxemburg, Norway, Germany, the Netherlands, Great Britain and France have higher intercepts for d12 than most other countries. This indicates that, for a given score on latent variable CONDITION, respondents from these countries judge it more important for immigrants to speak the country's official language. Apparently, this greater sensitivity for language as a condition for immigration is predominantly present in Western and Northern Europe. In Slovenia, the opposite pattern can be observed.

The final model has an acceptable fit. Clearly, partial scalar equivalence does not hold for all countries. The largest possible subset for which CONDITION possesses the characteristic of partial scalar equivalence is found when items d10 and d16 are chosen to calibrate the scale. But even then, the validity of latent country means is not guaranteed in seven countries (Czech Republic, Denmark, Finland, Hungary, Ireland, Italy and Portugal).

3.3 Metric and scalar equivalence tests for ECOTHREAT

Of the four immigration scales, perceived economic threat turns out to be the most problematic one with respect to cross-cultural equivalence. Fit indices indicate that the fit of the configural equivalence model is not splendid but acceptable. Imposing full metric equivalence, however, causes a dramatic increase in the chi-square value. No less than 14 substantial deviations from full metric equivalence were detected. Items d21 and 27 are the greatest sources of inequivalence, with 4 violations each.

In Portugal, Greece and Poland, a lower factor loading is found for the item referring to long-term unemployed immigrants (d21). In these countries, the perception of economic threat is less centered around immigrants receiving unemployment benefits. In France, on the other hand, the belief that long-term unemployed immigrants should be made to leave is more strongly connected to economic threat.

The largest subset of countries for which partial metric equivalence holds is found when items d19 and d25 are chosen as the items calibrating the scale. In this case, comparisons of regression coefficients are dubious for three countries, namely Czech Republic, Ireland and Italy.

After imposing scalar equivalence, the model could be improved substantially on 22 additional points. Seven of the violations of scalar equivalence relate to the intercepts of the item questioning whether immigrants contribute more to taxes than they take out. Here, a clear divide between Northern countries on one side,

and Southern and Eastern countries on the other can be discerned. Controlling for their score on ECOTHREAT, respondents from Sweden and Denmark—the two countries with the highest social expenditure (relative to GDP)[7]—are more inclined to answer that immigrants take more out of the welfare system than they put in. In countries with a less expansive welfare system such as Portugal, Spain, Italy, Slovenia and Poland, the concern that immigrants do not contribute proportionally to taxes is less widespread among the population.

Due to the large number of deviating parameters, partial scalar equivalence can be established for a relatively small number of countries only. Taking d19 and d25 as calibrating items again, cross-national mean comparisons are questionable for seven out of 21 countries. These countries are Belgium, Czech Republic, Spain, Greece, Hungary, Ireland and Italy.

Table 4 Equivalence tests for ECOTHREAT – N = 33,855

Model specifications		χ^2	df	RMSEA	CFI	$\Delta\chi^2$	EPC
M0	Configural equivalence	910.72	105	0.069	0.97	- -	- -
M1	Full metric invariance	1740.04	185	0.072	0.94	- -	- -
M2	λ_{d21}^{FR} free	1635.58	184	0.070	0.94	104.46	0.16
M3	λ_{d27}^{DE} free	1585.74	183	0.069	0.94	49.84	-0.15
M4	λ_{d27}^{GR} free	1540.88	182	0.068	0.95	44.86	-0.13
M5	λ_{d27}^{AT} free	1490.52	181	0.067	0.95	50.36	-0.16
M6	λ_{d19}^{IE} free	1459.23	180	0.066	0.95	31.29	0.13
M7	λ_{d26}^{FI} free	1426.15	179	0.066	0.95	33.08	-0.17
M8	λ_{d27}^{DK} free	1394.63	178	0.065	0.95	31.52	0.14
M9	λ_{d21}^{PT} free	1366.44	177	0.065	0.95	28.19	-0.18
M10	λ_{d26}^{PT} free	1334.95	176	0.064	0.95	31.49	-0.19
M11	λ_{d21}^{GR} free	1311.34	175	0.063	0.95	23.61	-0.10
M12	λ_{d26}^{GR} free	1275.14	174	0.063	0.96	36.20	-0.12
M13	λ_{d21}^{PL} free	1248.27	173	0.062	0.96	26.87	-0.13
M14	λ_{d19}^{CZ} free	1219.75	172	0.061	0.96	28.52	0.12
M15	λ_{d25}^{IT} free	1196.75	171	0.061	0.96	23.00	-0.26
M16	Scalar equivalence	4217.11	251	0.099	0.84	- -	- -
M17	τ_{d26}^{PT} free	3856.36	250	0.095	0.86	360.75	0.65
M18	τ_{d27}^{SE} free	3620.29	249	0.092	0.87	236.07	-0.55
M19	τ_{d27}^{DK} free	3287.07	248	0.087	0.88	333.22	-0.65
M20	τ_{d19}^{GR} free	3113.74	247	0.085	0.89	173.33	-0.41
M21	τ_{d26}^{DK} free	2995.37	246	0.083	0.89	118.37	-0.45

M22	τ_{d26}^{SE} free	2846.78	245	0.081	0.9	148.59	-0.43
M23	τ_{d19}^{HU} free	2683.34	244	0.079	0.9	163.44	-0.37
M24	τ_{d27}^{LU} free	2579.11	243	0.077	0.91	104.23	0.5
M25	τ_{d21}^{CZ} free	2499.63	242	0.076	0.91	79.48	-0.36
M26	τ_{d21}^{ES} free	2418.33	241	0.075	0.91	81.30	0.31
M27	τ_{d27}^{NO} free	2319.72	240	0.073	0.92	98.61	-0.32
M28	τ_{d21}^{FR} free	2250.00	239	0.072	0.92	69.72	0.41
M29	τ_{d26}^{IT} free	2178.38	238	0.071	0.92	71.62	0.35
M30	λ_{d27}^{SE} free	2124.41	237	0.07	0.92	53.97	0.11
M31	τ_{d27}^{NL} free	2072.76	236	0.07	0.93	51.65	-0.27
M32	τ_{d19}^{BE} free	2007.32	235	0.068	0.93	65.44	0.23
M33	τ_{d21}^{HU} free	1950.38	234	0.067	0.93	56.94	-0.25
M34	τ_{d26}^{SI} free	1897.38	233	0.067	0.93	53.00	0.28
M35	τ_{d26}^{ES} free	1850.17	232	0.066	0.94	47.21	0.25
M36	τ_{d27}^{PT} free	1809.60	231	0.065	0.94	40.57	0.32
M37	τ_{d26}^{PL} free	1760.92	230	0.064	0.94	48.68	0.25
M38	τ_{d21}^{IT} free	1737.20	229	0.064	0.94	23.72	-0.24
M39	τ_{d21}^{GR} free	1709.49	228	0.064	0.94	27.71	-0.23
M40	λ_{d25}^{GB} free	1684.55	227	0.063	0.94	24.94	-0.1

Common solution for the final model

λ_{d19}	λ_{d21}	$^*\lambda_{d25}$	λ_{d26}	λ_{d27}	τ_{d19}	τ_{d21}	$^*\tau_{d25}$	τ_{d26}	τ_{d27}
0.79	0.81	1.00	0.99	1.13	0.14	0.12	0.15	0.08	0.23

* Marker item (loading fixed to 1 to identify the model)

3.4 Metric and scalar equivalence tests for CULTHREAT

Fourth and last, cross-cultural equivalence is tested for the scale measuring perceived cultural treath (CULTHREAT).[8] The configural equivalent model is just identified and therefore has a perfect fit. Constraining factor loadings to be equal across countries does not cause too much misfit. The metric equivalent model can only be improved substantially on two points. After these modifications, RMSEA is well below 0.05. CFI equals 0.99 and does not depart too much from the CFI in the baseline model. Despite the small number of violations of the equivalence hypotheses, partial equivalence does not hold for all countries. One country should be excluded from comparing regression coefficients, depending on the two calibration items that are chosen.

Scalar equivalence poses more serious problems. Fourteen additional equality constraints cause substantial misfit and where therefore deleted. Six countries have

a deviating intercept for item d41, referring to the desirability of having a variety of religions in the country. In Austria, Germany, Denmark and the Netherlands, a higher intercept is found. Respondents from these countries reject[9] religious diversity more often than what can be expected from their levels of perceived cultural threat. In France and Hungary, the reverse pattern is found.

Also here, latent mean comparisons are problematical for a quite large number of countries. If d28 and d40 are chosen as calibration items, partial scalar equivalence does not hold for Switzerland, Spain, Finland, Great Britain, Ireland, Poland, Portugal and Sweden.

Table 5 Equivalence tests for CULTHREAT – N = 33,855

Model specifications		χ^2	Df	RMSEA	CFI	$\Delta\chi^2$	EPC
M0	Configural equivalence	0.00	0	0	1	- -	- -
M1	Full metric invariance	205.31	38	0.052	0.98	- -	- -
M2	λ_{d41}^{NO} free	164.53	37	0.046	0.99	40.78	0.25
M3	λ_{d40}^{CH} free	139.51	36	0.042	0.99	25.02	-0.27
M4	Scalar equivalence	2307.91	74	0.135	0.78	- -	- -
M5	τ_{d28}^{FI} free	2017.31	73	0.127	0.8	290.60	0.63
M6	τ_{d28}^{PL} free	1678.42	72	0.116	0.84	338.89	0.57
M7	τ_{d41}^{NL} free	1335.60	71	0.104	0.87	342.82	0.64
M8	τ_{d41}^{FR} free	1064.43	70	0.093	0.9	271.17	-0.65
M9	τ_{d28}^{SE} free	823.94	69	0.081	0.92	240.49	0.60
M10	τ_{d41}^{HU} free	723.41	68	0.076	0.93	100.53	-0.37
M11	τ_{d28}^{GB} free	623.69	67	0.071	0.94	99.72	-0.36
M12	τ_{d41}^{DE} free	540.68	66	0.066	0.95	83.01	0.27
M13	λ_{d28}^{PT} free	477.81	65	0.062	0.96	62.87	-0.33
M14	τ_{d41}^{AT} free	408.94	64	0.057	0.96	68.87	0.21
M15	τ_{d40}^{SE} free	370.83	63	0.054	0.97	38.11	0.39
M16	τ_{d28}^{IE} free	337.15	62	0.052	0.97	33.68	-0.2
M17	τ_{d41}^{DK} free	308.46	61	0.05	0.98	28.69	0.25
M18	τ_{d40}^{ES} free	278.60	60	0.047	0.98	29.86	-0.23

Common solution for the final model

λ_{d28}	$^*\lambda_{d40}$	λ_{d41}	τ_{d28}	$^*\tau_{d40}$	τ_{d41}
0.89	1.00	-0.87	0.12	0.22	-0.27

* Marker item (loading fixed to 1 to identify the model)

4. CONCLUSION AND DISCUSSION

Comparing abstract psychological constructs across cultures can be a difficult task. After all, it is not sure whether the items used to measure the constructs mean the same thing to members of different groups. Before valid comparisons can be made, equivalence of the measurement scales needs to be assessed. Neglecting this necessary condition for making cross-cultural comparisons can have serious consequences. Observed cross-cultural differences due to divergent question wording or differential functioning of the item might be erroneously interpreted in terms of real differences. Conversely, substantive differences might be obscured by cross-cultural differences in the interpretation of the items.

In this contribution, we evaluated measurement equivalence of four ESS-scales that indicate various aspects of attitudes toward immigration: opposition to allowing immigrants into the country (REJECT), support for conditions for immigration (CONDITION), perceived economic threat (ECOTHREAT) and perceived cultural threat (CULTHREAT). To test for measurement equivalence, a MGCFA approach was adopted. Various hierarchically ordered levels of equivalence were assessed. Configural equivalence implies that the same structure of salient and non-salient loadings holds in all countries under study. This level of equivalence, however, does not guarantee any score comparability. Metric equivalence presupposes the cross-national equality of the factor loadings, and renders it possible to compare regression coefficients and covariances in a valid way across countries. If not only factor loadings but also intercepts are equal across countries, then scalar equivalence is evidenced. Scalar equivalence is a necessary and sufficient condition for comparing country-means (Steenkamp & Baumgartner 1998). We argued that it is not necessary for all items to be measured invariantly. Meaningful comparisons can be made if equivalence holds for at least two items per construct. The latter situation is called partial equivalence (Byrne et al. 1989).

The results of the equivalence tests are summarized in Table 6. The degree of cross-cultural equivalence differs quite strongly from one scale to another. For REJECT, the number of violated equality constraints is limited, and partial scalar equivalence is found to hold for all countries. As a result, cross-country mean comparisons for this scale are warranted The high level of cross-cultural comparability of the REJECT-scale can probably be attributed to its rather abstract content. The conclusions are quite different for the other three scales, i.e. CONDITION, ECOTHREAT and CULTHREAT. Since partial metric equivalence is found to hold for 18 to 19 countries out of 21, the cross-country comparability of regression coefficients and covariances is guaranteed not for all, but for most countries. Partial scalar equivalence is found to be more problematic, and

meaningful country-mean comparisons are guaranteed for 11 (CULTHREAT) to 14 (CONDITION, ECOTHREAT) countries only.

Table 6 Summary of results

	REJECT	CONDITION	ECOTHREAT	CULTHREAT
Configural equivalence	all countries	all countries	all countries	all countries
(Partial) metric equivalence: *possible to compare regression coefficients*	all countries	18 countries	18 countries	19 countries
(Partial) scalar equivalence: *possible to compare country-means*	all countries	14 countries	14 countries	11 countries

A crucial question is what one should do when confronted with measurement inequivalence. In our opinion, a lack of measurement equivalence should certainly not lead to precluding all substantive analysis. Even when only configural equivalence is found, one can still perform separate analyses per country, and look for similarities or divergences in broad patterns of relations (although one should keep in mind that such a comparison has a more qualitative character, and that no parameters, such as effect sizes or country means, should be compared directly). Furthermore, the literature suggests various strategies to deal with deviations from equivalence (see, for example, Poortinga 1989). One can try to reduce the inequivalence by dropping certain countries or scales from the analysis. An alternative strategy is to treat measurement as a source of useful information on cross-cultural differences. In this study, for example, we found that Hungarians –more than inhabitants from other countries- interpret the term immigrants as referring to persons from a different ethnic group. Detection of inequivalence is thus not a finishing point, but a challenge for further research using additional data that are not included in the surveys.

Above all, this contribution has shown the importance of assessing measurement equivalence. Equivalence testing should become a standard practice for cross-cultural survey researchers. The practical research strategy offered in this paper can serve as a guideline hereby.

NOTES

1 These items are d4, d5, d9, d27, d28 and d29.
2 Apart from the these European countries, also Israel participated in ESS round 1. However, we decided to exclude Israel because of the large differences in the context of immigration and ethnic minorities in this country.
3 The EM-algorithm implemented in PRELIS was used to impute values for respondents who had a limited number of missing values (i.e. less than half of the items per scale).
4 In fact, ESS round 1 contains more items designed to measure the four theoretical concepts than the items mentioned here. However, preliminary analysis pointed out that several items do not measure the theoretical constructs in a clear-cut manner. Consequently, we decided to drop these items. Concretely, items were excluded if standardized factor loadings were too low (<.40) in a large number of countries (this is the case for items d14, d15, d20, d42, d43, d50, d53) or if modification indices suggested very strong error correlations that are theoretically not justified (items d9, d13, d17, d18, d44).
5 Factor loadings for the 21 countries are not given here, but can be obtained from the first author.
6 Correlations for the respective countries are not included in this paper, but can be requested from the first author.
7 Based on Eurostat figures on social expenditure. For more information, see the Eurostat web site: http://ec.europa.eu/eurostat.
8 Luxemburg is excluded from this analysis because of a non-positive definite input matrix. It is a well-known problem that polychoric correlation matrices turn out to be non-positive definite, since polychoric correlation matrices are not estimated as a whole, but each correlation coefficient is instead calculated based on a separate model (Wothke 1993). In this case, estimation problems – such as extremely large standard errors and negative variance estimates– occurred.
9 A higher score on this item expresses a stronger disagreement with the statement that a variety of religious is desirable.

REFERENCES

Bail, C. A. 2008. The Configuration of Symbolic Boundaries Against Immigrants in Europe. *American Sociological Review* 73: 37–59.

Bentler, P. M. 1990. Comparative Fit Indexes in Structural Models. *Psychological Bulletin* 107: 238–246.

Billiet, J. 2003. Cross-cultural Equivalence with Structural Equation Modeling. Harkness, J. A., Van de Vijver, F. J. R. & Mohler, P. Ph. (eds). *Cross-cultural survey methods* (pp. 247–264). Hoboken (N.J.): John Wiley & Sons.

Bollen, K. A. 1989. *Structural Equations with Latent Variables*. New York: John Wiley & Sons.

Brown, T.A. 2006. *Confirmatory Factor Analysis for Applied Research.* London: The Guilford Press.

Browne, M. W., & Cudeck, R. 1992. Alternative Ways of Assessing Model Fit. *Sociological Methods & Research* 21(2): 230–258.

Byrne, B. 1998. *Structural Equation Modeling wiht LISREL, PRELIS and SIMPLIS: Basic Concepts, Applications and Programming.* Mahwah (NJ): Lawrence Erlbaum.

Byrne, B. M., Shavelson, R. J., & Muthén, B. 1989. Testing for the Equivalence of Factor Covariance and Mean Structures: The Issue of Partial Measurement Invariance. *Psychological Bulletin* 105: 456–466.

Ceobanu A. M. & Escandell, X. 2010. Comparative Analyses of Public Attitudes Toward Immigrants and Immigration Using Multinational Survey Data: A review of Theories and Research. *Annual Review of Sociology* 36: 309–328.

Cheung, G. W., & Rensvold, R. B. 1999. Testing Factorial Invariance Across Groups: A Reconceptualization and Proposed New Method. *Journal of Management* 25(1): 1–27.

Davidov, E. & Meuleman, B. 2012. Explaining Attitudes Towards Immigration Policies in European Countries: The Role of Human Values. *Journal of Ethnic and Migration Studies* 38(5): 757–775.

Davidov, E., Datler, G., Schmidt, P. & Schwartz, S.H. 2011. Testing the Invariance of Values in the Benelux Countries With the European Social Survey: Accounting for Ordinality. In Davidov, E., Schmidt, P. & Billiet, J. (eds). *Cross-cultural Analysis: Methods and applications* (pp. 149–168). Routledge: London.

Drasgow, F., & Kanfer, R. 1985. Equivalence of Psychological Measurement in Heterogeneous Populations. *Journal of Applied Psychology* 70(4): 662–680.

Horn, J. L., & McArdle, J. J. 1992. A Practical and Theoretical Guide to Measurement Invariance in Aging Research. *Experimental Aging Research* 18(3): 117–144.

Hu, L., & Bentler, P. M. 1999. Cutoff Criteria for Fit Indexes in Covariance Structure Analysis: Conventional Criteria Versus New Alternatives. *Structural Equation Modeling* 6(1): 1–55.

Johnson, T. P. 1998. Approaches to Equivalence in Cross-cultural and Cross-national Survey-research. *ZUMA Nachrichten Spezial, Cross-Cultural Survey Equivalence* 3: 1–40.

Jöreskog, K. G. 1971. Simultaneous Factor Analysis in Several Populations. *Psychometrika,* 36(4): 408–426.

Jöreskog, K. G. 1990. New Developments in LISREL: Analysis of Ordinal Variables Using Polychoric Correlations and Weighted Least Squares. *Quality and Quantity* 24(4): 387–404.

Jöreskog, K. G., & Sörbom, D. 1993. *LISREL 8 User's Reference Guide.* Mooresville: Scientific Software.

Kankaras, M., Vermunt, J.K. & Moors, G. 2011. Measurement Rquivalence of Ordinal Items: A Comparison of Factor Analytic, Item Response Theory, and Latent Class Approaches. *Sociological Methods & Research* 40(2): 279–310.

Kunovich, R. M. 2004. Social Structural Position and Prejudice: An Exploration of Crossnational Differences in Regression Slopes. *Social Science Research* 33*:* 20–44.

Marsh, H. W. 1994. Confirmatory Factor Analysis Models of Factorial Invariance: A Multifaceted Approach. *Structural Equation Modeling* 1(1): 5–34.

Meuleman, B. 2012. When Are Item Intercept Differences Substantively Relevant in Measurement Invariance Testing?. In Salzborn S., Davidov E., Reinecke J. (eds). *Methods, Theories, and Empirical Applications in the Social Sciences. Festschrift fot Peter Schmidt* (pp. 97–104). Wiesbaden: Springer VS.

Meuleman, B., Davidov, E., & Billiet, J. 2009. Changing Attitudes Toward Immigration in Europe, 2002–2007: A Dynamic Group Conflict Theory Approach. *Social Science Research* 38(2): 352–365.

Millsap, R. E., & Yun-Tein, J. 2004. Assessing Factorial Invariance in Ordered-Categorical Measures. *Multivariate Behavioral Research* 39(3): 479–515.

Poortinga, Y. H. 1989. Equivalence of Cross-cultural Data: An Overview of Basic Issues. *International Journal of Psychology* 24: 737–756.

Reise, S. P., Widaman, K. F., & Pugh, R. H. 1993. Confirmatory Factor Analysis and Item Response Theory: Two Approaches for Exploring Measurement Invariance. *Psychological Bulletin* 114(3): 552–566.

Rensvold, R. B., & Cheung, G. W. 1998. Testing Measurement Models for Factorial Invariance: A Systematic Approach. *Educational and Psychological Measurement* 58: 1017–1034.

Saris, W. E., Satorra, A., & Sörbom, D. 1987. The Detection and Correction of Specification Errors in Structural Equation Models. *Sociological Methodology* 17: 105–129.

Schlueter, E. & Scheepers, P. 2010. The Relationship between Outgroup Size and Anti-Outgroup Attitudes: A Theoretical Synthesis and Empirical Test of Group Threat and Intergroup Contact Theory. *Social Science Research* 39: 285–295.

Schneider, S.L. 2008. Anti-immigrant Attitudes in Europe: Outgroup Size and Perceived Ethnic Threat. *European Sociological Review* 24(1): 53–67.

Semyonov, M., Raijman, R., & Gorodzeisky, A. 2006. The Rise of Anti-foreigner Sentiment in European Societies, 1988–2000. *American Sociological Review* 71: 426–449.

Sides, J., & Citrin, J. 2007. European Opinion About Immigration: The Role of Identities, Interests and Information. *British Journal of Political Science* 37: 477–504.

Steenkamp, J. E., & Baumgartner, H. 1998. Assessing Measurement Invariance in Cross-national Consumer Research. *Journal of Consumer Research* 25: 78–90.

Stephan, W.G., Ybarra, O., Martínez Martínez, C., Schwarzwald, J., & Tur-Kaspa, M. 1998. Prejudice Toward Immigrants to Spain and Israel. An Integrated Threat Theory Analysis. *Journal of Cross-Cultural Psychology* 29(4): 559–576.

Strabac, Z., & Listhaug, O. 2008. Anti-Muslim Prejudice in Europe: A multi-level Analysis of Survey Data from 30 Countries. *Social Science Research* 37: 268–286.

Van de Vijver, F., & Leung, K. 1997. *Methods and Data-analysis for Cross-cultural Research*. London: Sage.

Vandenberg, R. J., & Lance, C. E. 2000. A Review and Synthesis of the Measurement Invariance Literature: Suggestions, Practices, and Recommendations for Organizational Research. *Organizational Research Methods* 3: 4–70.

Wothke, W. 1993. Nonpositive Definite Matrices in Structural Modeling. In Bollen, K. A. & Long, J. S. (eds). *Testing structural equation models* (pp. 256–293). Newbury Park (CA): Sage.

Bart Meuleman is an Assistant Professor at the Centre for Sociological Research (CeSO), University of Leuven (Belgium), where he teaches research methodology. His main research interests involve cross-cultural comparisons of attitude and value patterns, such as welfare attitudes, ethnocentrism, religiosity and basic human values. In his work he mainly applies multilevel and structural equation models. Recent articles by him have appeared in Journal of Cross-Cultural Psychology, Journal of European Social Policy, Social Science Research and International Journal of Social Welfare.

E-mail: bart.meuleman@soc.kuleuven.be

Jaak Billiet, PhD in the Social Sciences, was full professor in social methodology at the Katholieke Universiteit Leuven,Belgium, and is since 2007 emeritus professor. He is a member of the central co–ordination team of the European Social Survey. His main research interest in methodology deals with validity assessment, interviewer and response effects, and the modeling of measurement error in social surveys. His substantial research covers longitudinal and comparative research in the domains of ethnocentrism, political attitudes and religious orientations. He also plays a central role in the implementation of the fourth wave of the European Values Study in 2008.

E-mail: jaak.billiet@soc.kuleuven.be